Mini Project 01 - IMDB web scraping

```
library(tidyverse)
library(rvest) # scrape data from internet
```

```
Warning message in system("timedatectl", intern = TRUE):
“running command 'timedatectl' had status 1”
Warning message:
“Failed to locate timezone database”
── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 

✓ ggplot2 3.3.5     ✓ purrr   0.3.4
✓ tibble  3.1.5     ✓ dplyr   1.0.7
✓ tidyr   1.1.4     ✓ stringr 1.4.0
✓ readr   2.0.2     ✓ forcats 0.5.1

── Conflicts ────────────────────────────────────────── tidyverse_conflicts() 
✖ dplyr::filter()  masks stats::filter()
✖ purrr::flatten() masks jsonlite::flatten()
✖ dplyr::lag()     masks stats::lag()


Attaching package: ‘rvest’
```

```
url <- "https://www.imdb.com/search/title/?groups=top_100&sort=user_rating,desc"
```

```
print(url)
```

```
[1] "https://www.imdb.com/search/title/?groups=top_100&sort=user_rating,desc"
```

```
#read html
imdb <- read_html(url)
```

```
imdb
```

```
{html_document}
<html xmlns:og="http://ogp.me/ns#" xmlns:fb="http://www.facebook.com/2008/fbml"
[1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset=UTF-8 .
[2] <body id="styleguide-v2" class="fixed">\n            <img height="1" widt .
```

```
#movie title
titles <- imdb %>%
html_nodes("h3.lister-item-header") %>%
html_text2()
```

```
titles[1:10]
```

'1. The Shawshank Redemption (1994)' · '2. The Godfather (1972)' · '3. The Dark Knight (2008)' ·
'4. Schindler\'s List (1993)' · '5. The Lord of the Rings: The Return of the King (2003)' ·
'6. The Godfather Part II (1974)' · '7. 12 Angry Men (1957)' · '8. Pulp Fiction (1994)' ·
'9. The Lord of the Rings: The Fellowship of the Ring (2001)' · '10. Inception (2010)'

```
# rating
ratings <- imdb %>%
    html_nodes("div.ratings-imdb-rating") %>%
    html_text2() %>%
    as.numeric()
```

```
ratings[1:10]
```

9.3 · 9.2 · 9 · 9 · 9 · 9 · 9 · 8.9 · 8.8 · 8.8

```
#number of vote
num_votes <- imdb %>%
    html_nodes("p.sort-num_votes-visible") %>%
    html_text2()
```

```
#build a dataframe
df <- data.frame(
        title = title,
        rating = ratings,
        num_vote = num_votes
)

head(df)
```

A data.frame: 6 × 3

	title	rating	num_vote
	<chr>	<dbl>	<chr>
1	1. The Shawshank Redemption (1994)	9.3	Votes: 2,707,341 | Gross: $28.34M | Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 1,880,128 | Gross: $134.97M | Top 250: #2
3	3. The Dark Knight (2008)	9.0	Votes: 2,680,664 | Gross: $534.86M | Top 250: #3
4	4. Schindler's List (1993)	9.0	Votes: 1,368,111 | Gross: $96.90M | Top 250: #6
5	5. The Lord of the Rings: The Return of the King (2003)	9.0	Votes: 1,863,607 | Gross: $377.85M | Top 250: #7
6	6. The Godfather Part II (1974)	9.0	Votes: 1,283,767 | Gross: $57.30M | Top 250: #4

Mini Project 02 - Specphone database

```
library(tidyverse)
library(rvest) # scrape data from internet
```

```
Warning message in system("timedatectl", intern = TRUE):
“running command 'timedatectl' had status 1”
Warning message:
“Failed to locate timezone database”
── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1

✓ ggplot2 3.3.5     ✓ purrr   0.3.4
✓ tibble  3.1.5     ✓ dplyr   1.0.7
✓ tidyr   1.1.4     ✓ stringr 1.4.0
✓ readr   2.0.2     ✓ forcats 0.5.1
```

```
── Conflicts ──────────────────────────────────────── tidyverse_conflicts()
✖ dplyr::filter()  masks stats::filter()
✖ purrr::flatten() masks jsonlite::flatten()
✖ dplyr::lag()     masks stats::lag()


Attaching package: ‘rvest’
```

```
url <- read_html("https://specphone.com/Apple-iPad-Mini-Wifi.html")
```

```
att<- url %>%
    html_nodes("div.topic") %>%
    html_text2()
value <- url %>%
    html_nodes("div.detail") %>%
    html_text2()
```

```
data.frame(attribute = att , value = value)
```

A data.frame: 32 × 2

attribute	value
<chr>	<chr>
วันเปิดตัว	**พฤศจิกายน 2555**
วันวางจำหน่าย	**พฤษภาคม 2556, สินค้าจำหน่ายหมดแล้ว**
ขนาด	200.00 x 134.70 x 7.20 มม.
น้ำหนัก	308 กรัม
วัสดุ	Aluminium, Plastic
SIM	**รองรับ 1 ซิมการ์ด**
Technology	EDGE, HSPA, HSPA+, LTE
2G	-
3G	-
4G	-
5G	-
ความเร็ว	EDGE, HSPA, HSPA+, LTE
ประเภท	IPS LCD
ขนาดหน้าจอ	7.90 นิ้ว
ความละเอียด	1024 x 768 pixels
ฟีเจอร์เพิ่มเติม	**โอนหรือย้ายข้อมูลบน PC ผ่าน iTunes มาพร้อมกับระบบปฏิบัติการ iOS 6.0 ป้องกันรอยนิ้วมือบนหน้าจอ (Oleophobic coating) ใช้กระจกแบบ Gorilla Glass ป้องกันรอยขีดข่วน**
ระบบปฏิบัติการ	iOS 6.00
ชิปประมวลผล	Apple A5 (APL2498) 1 GHz
ชิปกราฟิก	PowerVR SGX 543MP2
หน่วยความจำ	512 MB
ความจุ	16 GB
Memory Card	**ไม่รองรับ**
กล้องหลัก	**ตัวที่ 1: 5 MP**
ความละเอียดวีดีโอ	1080p (30 fps)
กล้องหน้า	**ตัวที่ 1: 1 MP**
Bluetooth	4
Wi-Fi	802.11b/g/n
USB	Lightning
GPS	GPS, AGPS
NFC	**ไม่รองรับ**
ความจุ	4,490 mAh
ประเภท	Li - Polymer

```
apple_url <- read_html("https://specphone.com/brand/Apple")
```

```
# link to all Apple
links <- apple_url %>%
    html_nodes("li.mobile-brand-item a") %>%
    html_attr("href")
```

```
full_links <- paste0("https://specphone.com",links)
```

```
result <- data.frame()

for ( link in full_links[1:10] ) {
    ap_topic <- link %>%
        read_html() %>%
        html_nodes("div.topic") %>%
        html_text2()

    ap_detail <- link %>%
        read_html() %>%
        html_nodes("div.detail") %>%
        html_text2()

    tmp <- data.frame(attribute = ap_topic,
                value = ap_detail)
    result <- bind_rows(result, tmp)
  }

print(result)
```

```
         attribute
1          วันเปิดตัว
2      วันวางจำหน่าย
3              ขนาด
4            น้ำหนัก
5              วัสดุ
6              SIM
7       Technology
8               2G
9               3G
10              4G
11              5G
12         ความเร็ว
13          ประเภท
14       ขนาดหน้าจอ
15       ความละเอียด
16     ฟีเจอร์เพิ่มเติม
17     ระบบปฏิบัติการ
18      ชิปประมวลผล
```

```
print(head(result),3)
```

```
    attribute                          value
1     วันเปิดตัว                 พฤศจิกายน 2555
2 วันวางจำหน่าย พฤษภาคม 2556, สินค้าจำหน่ายหมดแล้ว
3        ขนาด     200.00 x 134.70 x 7.20 มม.
4       น้ำหนัก                        308 กรัม
5         วัสดุ             Aluminium, Plastic
6         SIM                 รองรับ 1 ซิมการ์ด
```

```
#write csv
write_csv(result, "result_Apple.csv")
```